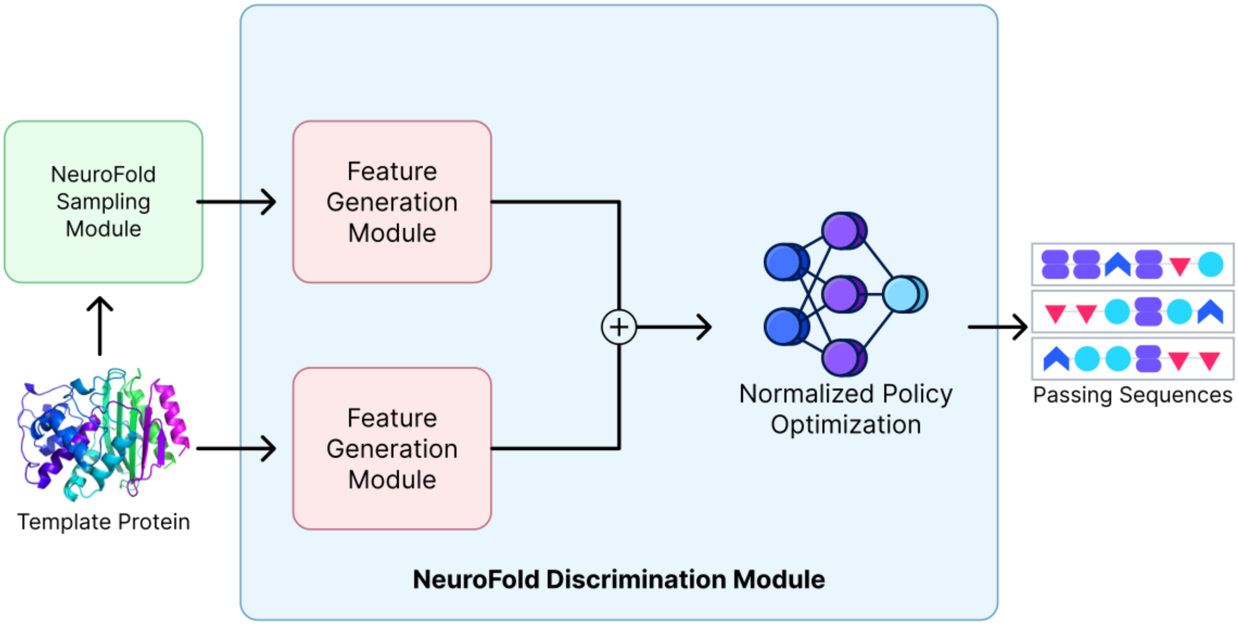
NeuroFold: A Multimodal Approach to Generating Novel Protein Variants *in silico*

Keaun Amani[1]*, Michael Fish[2]*, Matthew D. Smith[2], and Christian Danve M. Castroverde[2]

[1]Neurosnap Research, Neurosnap Inc., Wilmington, DE, USA; hello@neurosnap.ai (K.A.)
[2]Department of Biology, Wilfrid Laurier University, Waterloo, ON, CA; fish1960@mylaurier.ca (M.F.); msmith@wlu.ca (M.D.S.); dcastroverde@wlu.ca (C.D.M.C.)
*Corresponding Authors: K.A. and M.F.

Abstract

The generation of high-performance enzyme variants with desired physicochemical and functional properties presents a formidable challenge in the field of protein engineering. Existing *in silico* design methods are limited by inadequate training data, insufficient diversity within datasets, and suboptimal sampling techniques. Here, we introduce a novel approach that addresses these limitations and significantly improves the efficiency of generating functional enzyme variants. Using a multimodal approach, NeuroFold can leverage sequence, structural, and homology data during both sampling and discrimination phases, thereby enabling more diverse and informed sampling of the sequence space. Our model demonstrated a 40-fold increase in Spearman rank correlation as compared to large language models (LLMs) such as ESM-1v and empowers the rapid creation of high-quality enzyme variants, such as the β-lactamase variants generated by NeuroFold in this study, which demonstrated increased thermostability and varying levels of activity. This pipeline represents a promising advancement in the field of enzyme engineering, offering a valuable tool for the development of novel enzymes with enhanced performance and desired chemical properties.


NeuroFold Sampling Module
Feature Generation Module
Feature Generation Module
Template Protein
Normalized Policy Optimization
Passing Sequences
NeuroFold Discrimination Module

Introduction

Enzymes, as proteins that facilitate biochemical reactions, possess remarkable practical and industrial significance (Robinson, 2015). The ability to create customized enzyme variants with specific chemical properties, such as enhanced thermostability, increased or decreased reaction rate, broader pH stability, and improved solubility holds immense value in various applications (Li et al., 2020). However, generating such desired variants remains an exceptionally difficult task, primarily due to the magnitude of the protein sequence space (Romero & Arnold, 2009).

Conventional experimental methods, like deep mutational scans (DMS), prove undesirable for generating improved enzyme variants as they necessitate costly and highly parallelizable assays to be effective while only sampling a small portion of the protein sequence space, typically limited to a few point mutations or indels (Notin et al., 2022). For example, the fitness landscape of point mutations in three indole-3-glycerol phosphate synthase (IGPS) orthologs was assayed in one study to reveal the importance of both sequence and structural features of these model TIM barrels (Chan et al., 2017). Another study used multiplex assays to assess the differential abundance and activity of missense variants of the Vitamin K epoxide reductase (VKOR) enzyme (Chiasson et al., 2020). As demonstrated by these and other studies, the production of a substantial number of high-quality enzymes is severely restricted by the above-mentioned limitations, such as time-/labor-intensive assays and/or sub-universal protein sequence coverage.

Recently, the emergence of deep learning-based models has led to the development of novel *in silico* techniques for designing enzyme variants. Protein language models (pLMs) such as ESM-MSA (Rao et al., 2021), ESM-1v (Meier et al., 2021), and CARP-640m (Yang et al., 2022) have demonstrated varying levels of success in enzyme fitness prediction. These newer approaches offer the ability to generate better enzyme candidates more rapidly while streamlining the experimental bottleneck required in the past. Despite their advantages, deep learning models also have their limitations, including inconsistent accuracy and an inability to generalize effectively to certain proteins such as large and/or multimeric enzymes (Table 1).

Structural models such as AlphaFold2 are still unable to consistently predict mutations that impact protein structure (Table 1). Additionally, while single sequence methods like ESM-1v and CARP-640m have demonstrated reasonable accuracy for discriminating against mutations, they can sometimes struggle with orphaned sequences (proteins with low or no homology to known domains or proteins) (Table 1). We hypothesize that the limitations of ESM-1v, CARP-640m and related models are due to single sequence models being unable to capture co-evolutionary relationships that are not implicitly distilled into their model weights. This concept is reinforced further by evolutionary methods such as ESM-MSA, which demonstrate significantly greater performance when it comes to discriminating against mutations (Table 1). Arguably, this is due to being explicitly given an input multiple sequence alignment (MSA) from which to extract co-evolutionary relationships, as well as being able to implicitly capture information related to the protein landscape within their model weights.

In the current study, we present NeuroFold, an innovative hybrid model for designing enzyme variants with desired traits that incorporates a unique architecture to address the limitations of previous methods. Unlike conventional approaches that predominantly rely on a single modality such as sequence or evolutionary data in the form of an MSA (Rao et al., 2021), NeuroFold goes beyond reliance on a single modality by integrating sequence and evolutionary data with structural information. By combining these diverse data sources into a single multimodal approach, our model is able to better reason about the protein space, enabling the inclusion of theoretical constraints that guide the generation of viable enzymes. As a proof of principle, Neurofold was successfully used to design variants of β-lactamase with enhanced thermostability and varying levels of activity relative to the wild-type. As a result, NeuroFold achieves a broader coverage of the sequence space and demonstrates the potential for confidently producing enzyme variants with improved stability, desirable physicochemical properties, and enhanced reaction rates.

NeuroFold Architecture

Sampling Module

The NeuroFold architecture starts with an input amino acid sequence that acts as the template or reference. This template sequence should correspond to a functional enzyme that can also be used as a baseline for comparison (e.g., the wildtype version of an enzyme). This sequence can then be used to produce a large number of variants using either the MSA approach, an inverse folding approach, or a combination of the two methods (Figure 1).

The MSA approach consists of using ESM-MSA (Rao et al., 2021) with Gibbs sampling (Johnson et al., 2021) to produce novel sequences from an input MSA produced from the template sequence. This method is useful for sequences with rich homology and allows novel sequences to be inferred from evolutionary data (Rao et al., 2021).

The alternative inverse folding approach utilizes ProteinMPNN (Dauparas et al., 2022) to generate new sequences from a template structure. This is ideal when a high quality experimental structure is available for the template protein and was employed in this case (see Table S1 for settings). Inverse folding operates on the assumption that proteins with similar structures will tend to possess similar properties, an assumption that is supported by our findings.

The benefit of utilizing an inverse folding approach is that models such as ProteinMPNN are typically capable of sampling a much broader sequence space with sampled sequences, typically sharing 40-70% pairwise sequence identity to the wildtype. This can be a powerful advantage over MSA and traditional approaches such as DMS, as proteins with divergent sequences can still produce similar folds and even carry out similar functions. For example, Koehl and Levitt (2002) showed that the B1 domain of the streptococcal Protein G and *P. magnus* Protein L share striking structural similarities in spite of their low amino acid sequence identity (Figure S1).

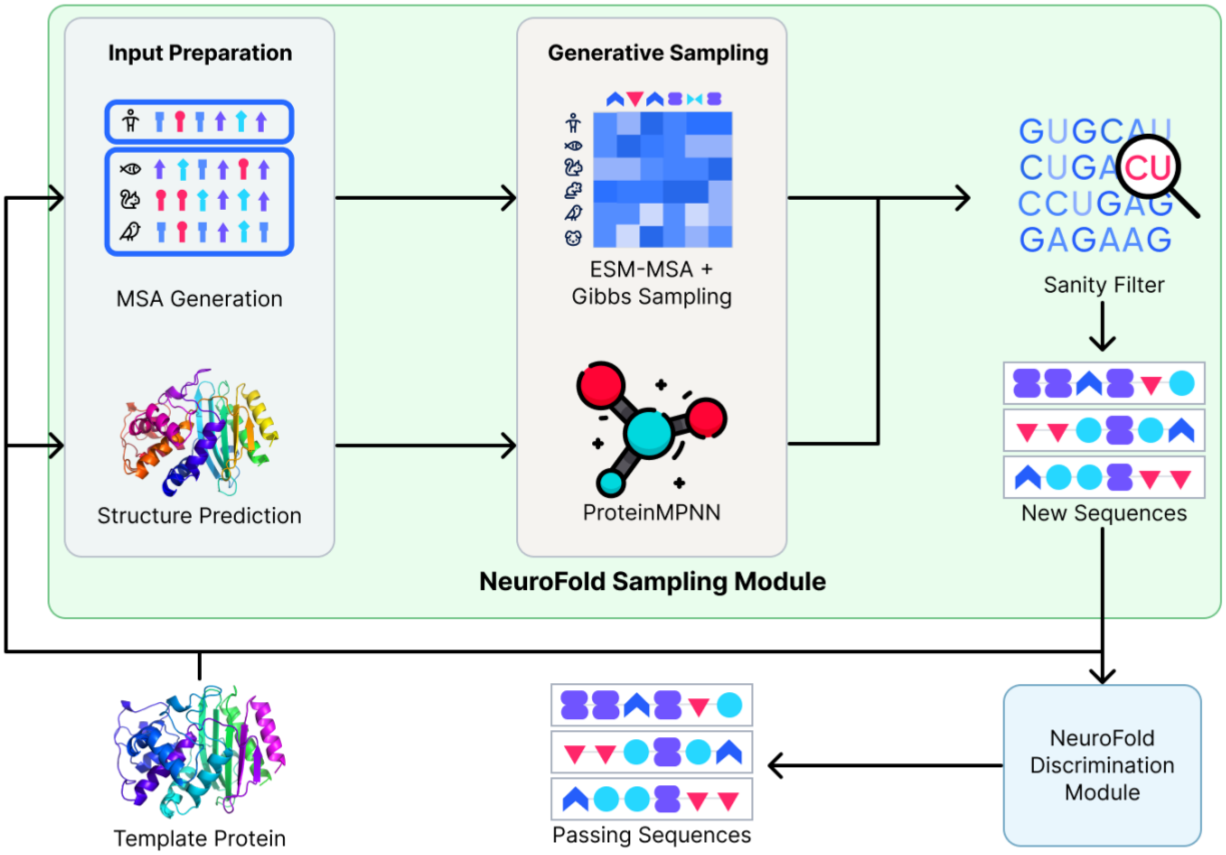


Figure 1. Outline of NeuroFold's Sampling Module. NeuroFold's architecture can be broken down into 3 core modules: The Sampling Module, Discrimination Module, and Feature Generation Module. The Sampling Module is responsible for producing the candidate sequences. The Discrimination Module is responsible for discriminating against functional and non-functional candidates generated by the sampling module. Finally, the Feature Generation Module is a component of the Discrimination Module that is responsible for producing the input features of each candidate sequence.

The sequences generated from the sampling models are then evaluated using a static sanity filter consisting of rules such as removing sequences that contain an excessive number of repeats as well as removing sequences that contain disruptive or rigid amino acids in loop regions. The new set containing the filtered sequences is then passed into the Discrimination Module alongside the original template sequence. The result is a subset of sequences predicted to be fit by the Discrimination Module. These sequences can then be experimentally validated or used for additional downstream evaluation.

Discrimination Module

The Discrimination Module (Figure 2) is responsible for determining whether or not a protein variant is fit with relation to a reference protein. This module receives two input sequences, a template sequence that acts as the reference and a mutated or variant sequence that acts as the candidate to evaluate. The template sequence is critical as it grounds the model with a baseline to compare as a reference. By introducing this baseline, the model learns to compare and measure the difference in potential fitness compared to just learning an arbitrary relative fitness for the entire protein landscape.

These sequences are then passed into the Feature Generation Module (Figure 3) to produce two sets of corresponding output representations. These output representations are then fed into a Normalized Policy Network (NPN) which classifies the mutated sequence as either fit or unfit relative to the template sequence. The NPN acts as a binary classifier using the output representation from the FGM as input.

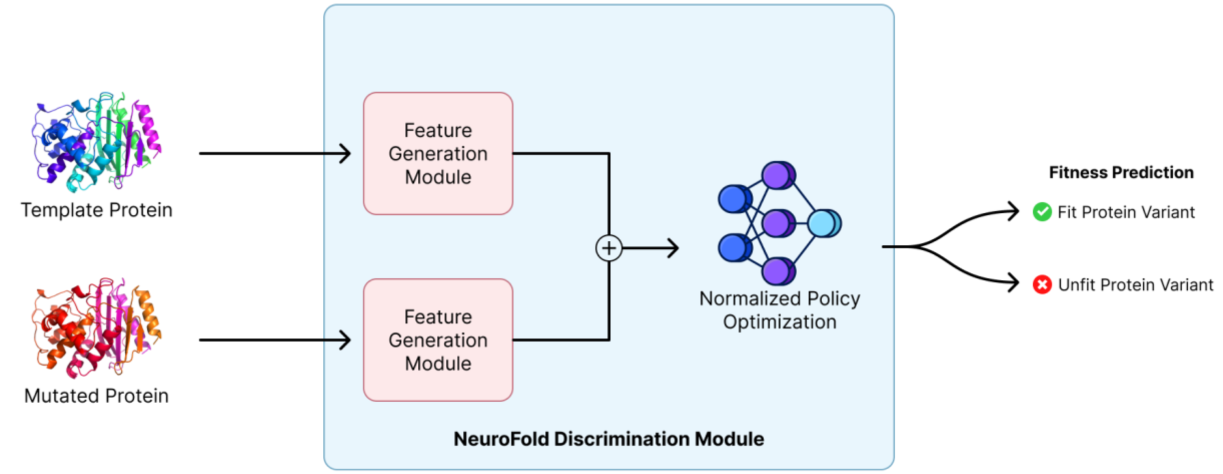

Template Protein
Mutated Protein
Feature Generation Module
Feature Generation Module
Normalized Policy Optimization
NeuroFold Discrimination Module
Fitness Prediction
Fit Protein Variant
Unfit Protein Variant

Figure 2. Outline of NeuroFold's Discrimination Module. The Discrimination Module is responsible for differentiating between a fit and unfit enzyme. The module itself effectively acts as a binary classifier that predicts whether an enzyme can be considered fit.

Feature Generation Module

The Feature Generation Module (FGM) is a key part of what gives NeuroFold its predictive power. The FGM receives an input amino acid sequence and generates a corresponding output representation which is used throughout the rest of the network. The FGM itself can be split into two main segments. The first segment is responsible for producing features from the input sequence in the form of an MSA and a predicted structure. These features are then passed into the Ensembling Module, a collection of models for each modality. The output from the models are then ensembled by modality and returned as an output representation (Figure 3).

The MSAs are generated locally using the MMseqs2 algorithm (Steinegger, 2017) with UniRef30_2022 as the database (Süzek et al., 2007; Mirdita et al., 2022). For each input sequence, two protein structures are predicted using AlphaFold2-ptm (Jumper et al., 2021; Mirdita et al., 2022); the first structure is predicted using the same generated MSA, while the other is predicted as a single sequence and without MSA as AlphaFold2 and RoseTTAFold (Baek et al., 2021) have both demonstrated higher accuracy on certain sequences when no MSA is provided (Baek & Baker, 2022). This enhanced representation biases the model to produce more realistic proteins that other models without the same constraints might ignore.

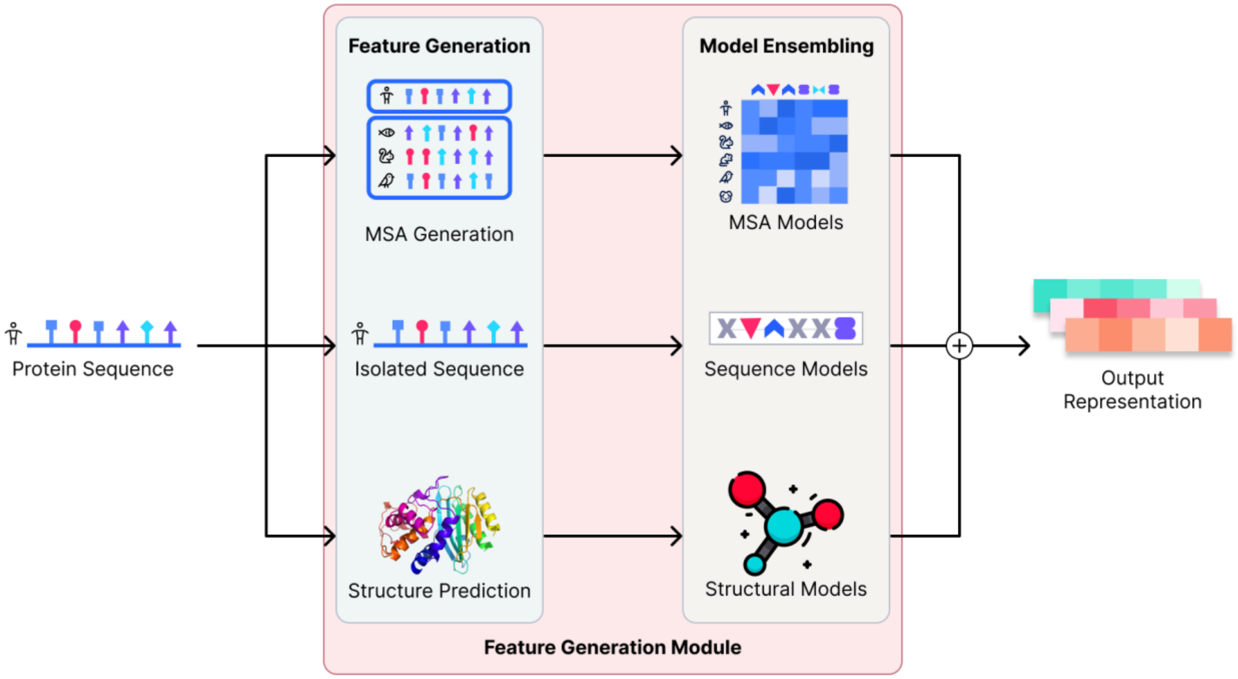


Figure 3. Diagrammatic representation of the NeuroFold Feature Generation Module (FGM). The FGM uses a multimodal approach to capture meaningful representations from sequence, evolutionary, and structural data.

Performance Metrics

NeuroFold was tested on experimentally validated proteins from the ProteinGym dataset (Notin et al., 2022), TAPE stability dataset (Rao, 2019; Rocklin et al., 2017), as well as additional datasets collected from various studies (Johnson et al., 2023; Madani et al., 2023; Repecka et al., 2021; Russ et al., 2020). The subset of proteins used from the ProteinGym dataset was carefully curated so as to exclude experimental data from proteins with irrelevant non-enzymatic functions in processes such as viral replication and protein-protein interactions.

The final dataset consisted of 116,321 individual protein sequences from 38 protein families/taxa (Johnson et al., 2023; Madani et al., 2023; Notin et al., 2022; Rao, 2019; Repecka et al., 2021; Rocklin et al., 2017; Russ et al., 2020), including variants with single point mutations, multiple point mutations, and indels. The benchmarks for fitness within these datasets consist of enzyme activity, enzyme stability, and fluorescence. In addition to NeuroFold testing, the collated experimental datasets were also fed into currently existing models, such as CARP-640m (Yang et al., 2022), ESM-1v (Meier et al., 2021), ESM-MSA (Rao et al., 2021), AlphaFold2 (Jumper et al., 2021), MIF-ST (Yang et al., 2022), ESM-IF (Hsu et al. 2022), ProteinMPNN (Dauparas et al., 2022), and SolubleMPNN.

NeuroFold outperformed all other models with the highest accuracy, f1-score and precision metrics, all of which were higher than alternative approaches (Table 1). For practical applications within the task of enzyme stability prediction, precision is the most valuable metric as it is not the number of false negatives that are of value, but the number of false positives. Since the protein space is so vast, the major limitation is rarely computationally sampling viable sequences. Instead, the bottleneck is typically the experimentation and validation of proteins. Minimizing false positives effectively minimizes the experimental time, cost, and resources.

NeuroFold demonstrated an accuracy of 70.25% when validated on our dataset (Table 1). The next best-performing model was ESM-MSA with 64.18%, with ESM-1v and single sequence AlphaFold2 scoring the lowest in this metric (48.06% and 44.37% respectively). Overall, the accuracy achieved by NeuroFold represents a significant improvement over the typical success rates of ~14.17% achieved using traditional methods (Notin et al., 2022; Table S2). The precision of NeuroFold on each dataset was also independently computed (Figure 4). We find the overall performance of AF2-SS to be significantly lower than that of AF2-MSA and arguably redundant. For this reason we would recommend excluding this method in the future.

Model performance on benchmark dataset.

Table 1. Evaluation of NeuroFold's performance using our dataset and comparison to various other models. Prediction accuracy, precision, recall, F1-score, and Spearman rank correlation with activity / fitness were calculated. The cutoffs used for all models was the sample 50th percentile. AF2-MSA corresponds to ColabFold with mmseqs2 MSA generation while AF2-SS corresponds to ColabFold with single sequence input.

Model	**Accuracy**	**Precision**	**Recall**	**F1-score**	**Spearman**
NeuroFold-base	**0.7025**	**0.7036**	**0.7443**	**0.7234**	**0.40**
CARP-640m	0.4812	0.5038	0.4820	0.4926	0.02
ESM-1v	0.4806	0.5032	0.4814	0.4921	0.01
ESM-1v mask6	0.4857	0.5083	0.4863	0.4970	0.01
ESM-MSA	0.6418	0.6644	0.6357	0.64.97	0.33
AF2-MSA pLDDT	0.5516	0.5742	0.5494	0.5615	0.08
AF2-MSA pTM	0.5565	0.5791	0.5541	0.5663	0.12
AF2-MSA PAE	0.5226	0.5452	0.5216	0.5331	0.01
AF2-SS pLDDT	0.4437	0.4663	0.4461	0.4560	0.14
AF2-SS pTM	0.4553	0.4779	0.4572	0.4673	0.10
AF2-SS PAE	0.4669	0.4895	0.4683	0.4786	-0.08
MIF-ST	0.5888	0.6114	0.5849	0.5979	0.18
ESM-IF	0.5544	0.5770	0.5520	0.5642	0.12
ProteinMPNN	0.5571	0.5797	0.5546	0.5669	0.14
SolubleMPNN	0.5556	0.5782	0.5532	0.5655	0.13

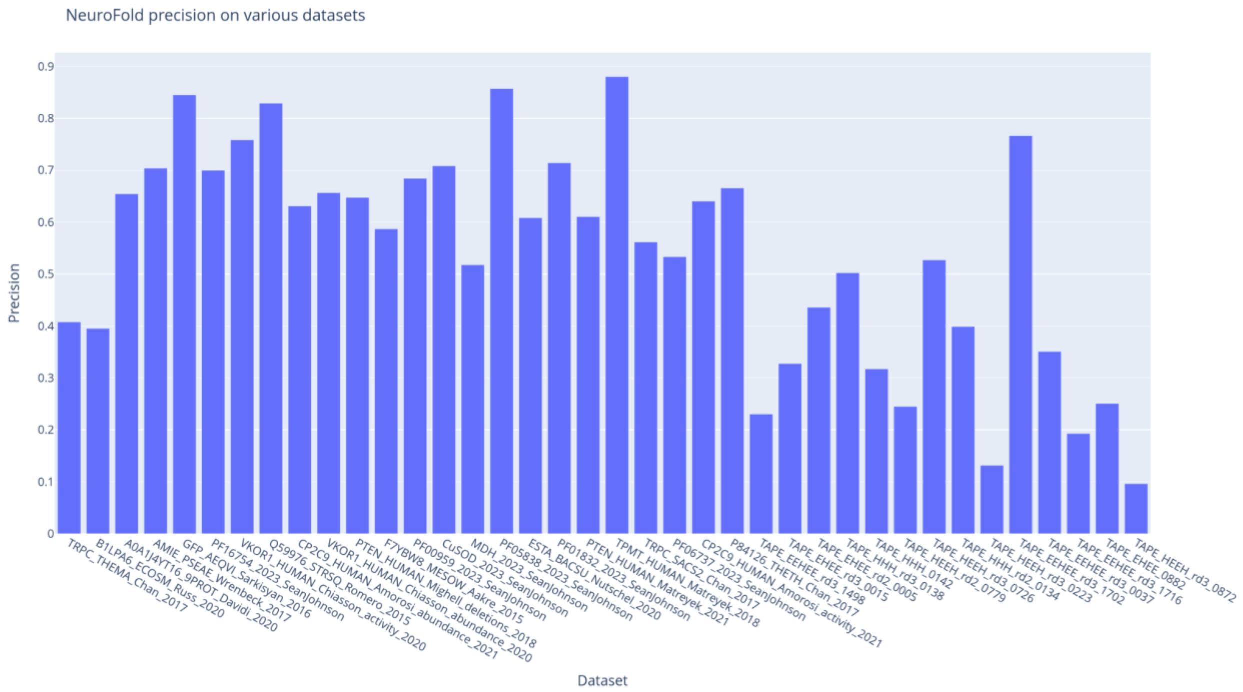
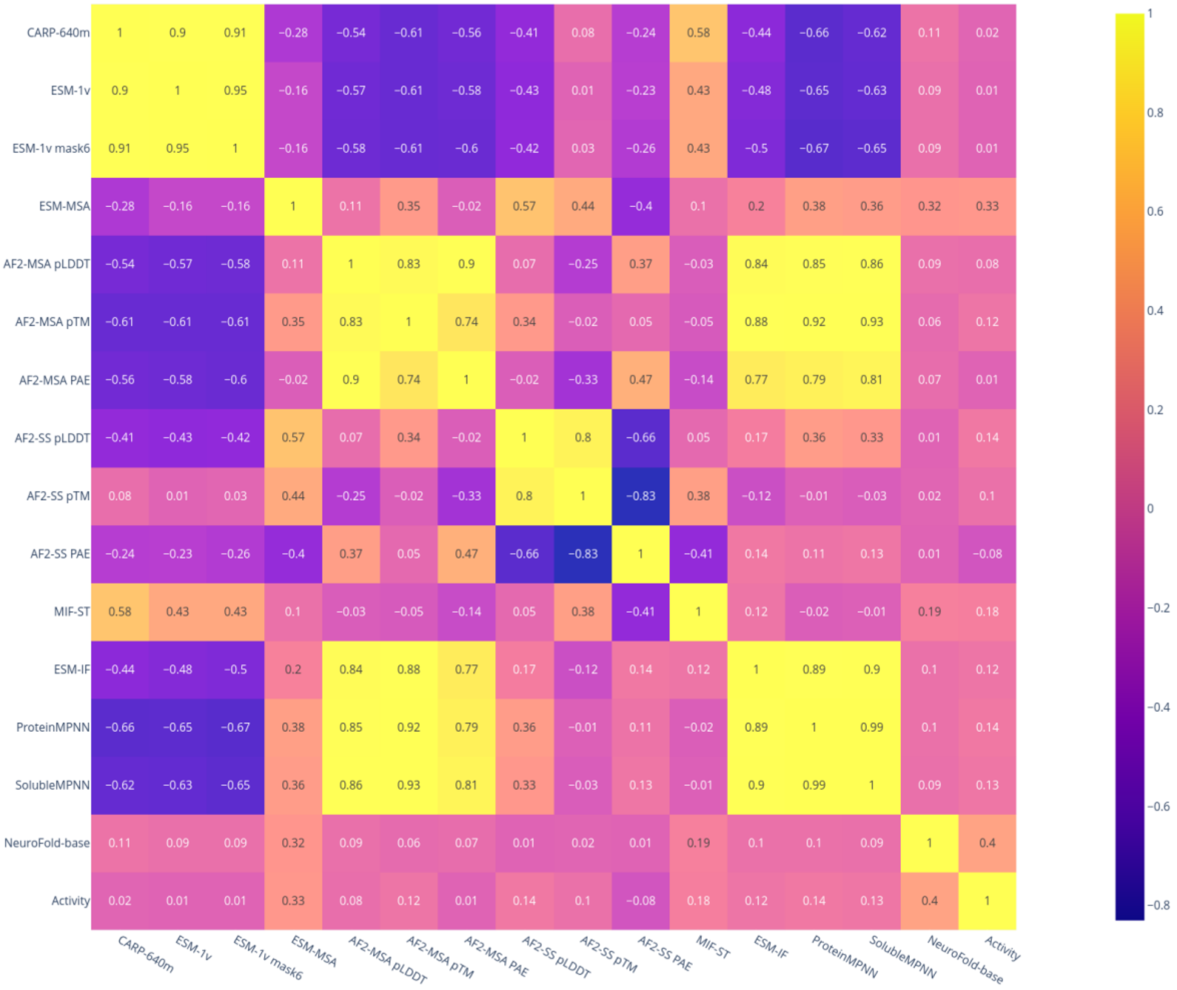


Figure 4. NeuroFold precision on various datasets. NeuroFold's precision was calculated on 38 different datasets of experimentally validated proteins. A high precision score indicates that NeuroFold correctly predicted the fitness of proteins tested within a specific dataset, while a low precision score indicates a smaller degree of correct predictions per protein variant. The details, computed metrics, computed MSAs, as well as all computed structures are all freely available to download from the following URL https://neurosnap.ai/neurofold/dataset.

We also benchmarked NeuroFold against various deep learning models using the Spearman rank correlation (Figure 5). The Spearman rank correlation with experimentally validated activity/fitness was benchmarked against each model's output on an input sequence. NeuroFold showed a 0.4 Spearman rank correlation with activity/fitness (Table 1). The next best model was ESM-MSA with a 0.33 Spearman rank correlation, which is a 17.5% decrease compared to NeuroFold. Interestingly, some models exhibited very low correlations (e.g. ESM-1v and AF2), further reflecting the strength of the NeuroFold architecture in robustly predicting protein fitness as determined by the collected experimental dataset. The Spearman correlation is another useful metric to use in conjunction with accuracy, and the aforementioned metrics as a broad criterion for evaluating a model's consistent ability to discriminate against unfit proteins (Notin et al., 2022). For the current study, however, we prioritized the optimization of precision and f1-score at the cost of correlation, as optimizing precision leads to more immediate practical applications. Overall, NeuroFold's performance appears to be superior on ProteinGym datasets compared to TAPE datasets. Performance is best on TPMT_HUMAN_Matreyek_2018 (Matreyek., 2018) and worst on TAPE_HEEH_rd3_0872 (Rocklin et al., 2017). It is unclear whether the low performance on certain datasets is due to experimental error, data curation error, or an inability of NeuroFold to generalize to those datasets.



	CARP-640m	ESM-1v	ESM-1v mask6	ESM-MSA	AF2-MSA pLDDT	AF2-MSA pTM	AF2-MSA PAE	AF2-SS pLDDT	AF2-SS pTM	AF2-SS PAE	MIF-ST	ESM-IF	ProteinMPNN	SolubleMPNN	NeuroFold-base	Activity
CARP-640m	1	0.9	0.91	−0.28	−0.54	−0.61	−0.56	−0.41	0.08	−0.24	0.58	−0.44	−0.66	−0.62	0.11	0.02
ESM-1v	0.9	1	0.95	−0.16	−0.57	−0.61	−0.58	−0.43	0.01	−0.23	0.43	−0.48	−0.65	−0.63	0.09	0.01
ESM-1v mask6	0.91	0.95	1	−0.16	−0.58	−0.61	−0.6	−0.42	0.03	−0.26	0.43	−0.5	−0.67	−0.65	0.09	0.01
ESM-MSA	−0.28	−0.16	−0.16	1	0.11	0.35	−0.02	0.57	0.44	−0.4	0.1	0.2	0.38	0.36	0.32	0.33
AF2-MSA pLDDT	−0.54	−0.57	−0.58	0.11	1	0.83	0.9	0.07	−0.25	0.37	−0.03	0.84	0.85	0.86	0.09	0.08
AF2-MSA pTM	−0.61	−0.61	−0.61	0.35	0.83	1	0.74	0.34	−0.02	0.05	−0.05	0.88	0.92	0.93	0.06	0.12
AF2-MSA PAE	−0.56	−0.58	−0.6	−0.02	0.9	0.74	1	−0.02	−0.33	0.47	−0.14	0.77	0.79	0.81	0.07	0.01
AF2-SS pLDDT	−0.41	−0.43	−0.42	0.57	0.07	0.34	−0.02	1	0.8	−0.66	0.05	0.17	0.36	0.33	0.01	0.14
AF2-SS pTM	0.08	0.01	0.03	0.44	−0.25	−0.02	−0.33	0.8	1	−0.83	0.38	−0.12	−0.01	−0.03	0.02	0.1
AF2-SS PAE	−0.24	−0.23	−0.26	−0.4	0.37	0.05	0.47	−0.66	−0.83	1	−0.41	0.14	0.11	0.13	0.01	−0.08
MIF-ST	0.58	0.43	0.43	0.1	−0.03	−0.05	−0.14	0.05	0.38	−0.41	1	0.12	−0.02	−0.01	0.19	0.18
ESM-IF	−0.44	−0.48	−0.5	0.2	0.84	0.88	0.77	0.17	−0.12	0.14	0.12	1	0.89	0.9	0.1	0.12
ProteinMPNN	−0.66	−0.65	−0.67	0.38	0.85	0.92	0.79	0.36	−0.01	0.11	−0.02	0.89	1	0.99	0.1	0.14
SolubleMPNN	−0.62	−0.63	−0.65	0.36	0.86	0.93	0.81	0.33	−0.03	0.13	−0.01	0.9	0.99	1	0.09	0.13
NeuroFold-base	0.11	0.09	0.09	0.32	0.09	0.06	0.07	0.01	0.02	0.01	0.19	0.1	0.1	0.09	1	0.4
Activity	0.02	0.01	0.01	0.33	0.08	0.12	0.01	0.14	0.1	−0.08	0.18	0.12	0.14	0.13	0.4	1

Figure 5. Spearman rank correlation matrix of diverse protein deep learning models. Different models (including NeuroFold) were tested on a collated dataset of 116,321 experimentally validated proteins from 38 protein families/taxa. Different outputs from several different models and algorithms were evaluated for their predictive power of protein fitness based on experimentally confirmed metrics. This correlation matrix highlights the Spearman rank correlation between all results and fitness. ESM-1v mask6 is ESM-1v (Meier et al., 2021) with every 6th amino acid masked, which has been demonstrated to improve discrimination ability (Johnson et al., 2023). SolubleMPNN is ProteinMPNN trained on a dataset of soluble proteins only.

Experimental Validation

Generation of β-Lactamase Variants Using NeuroFold

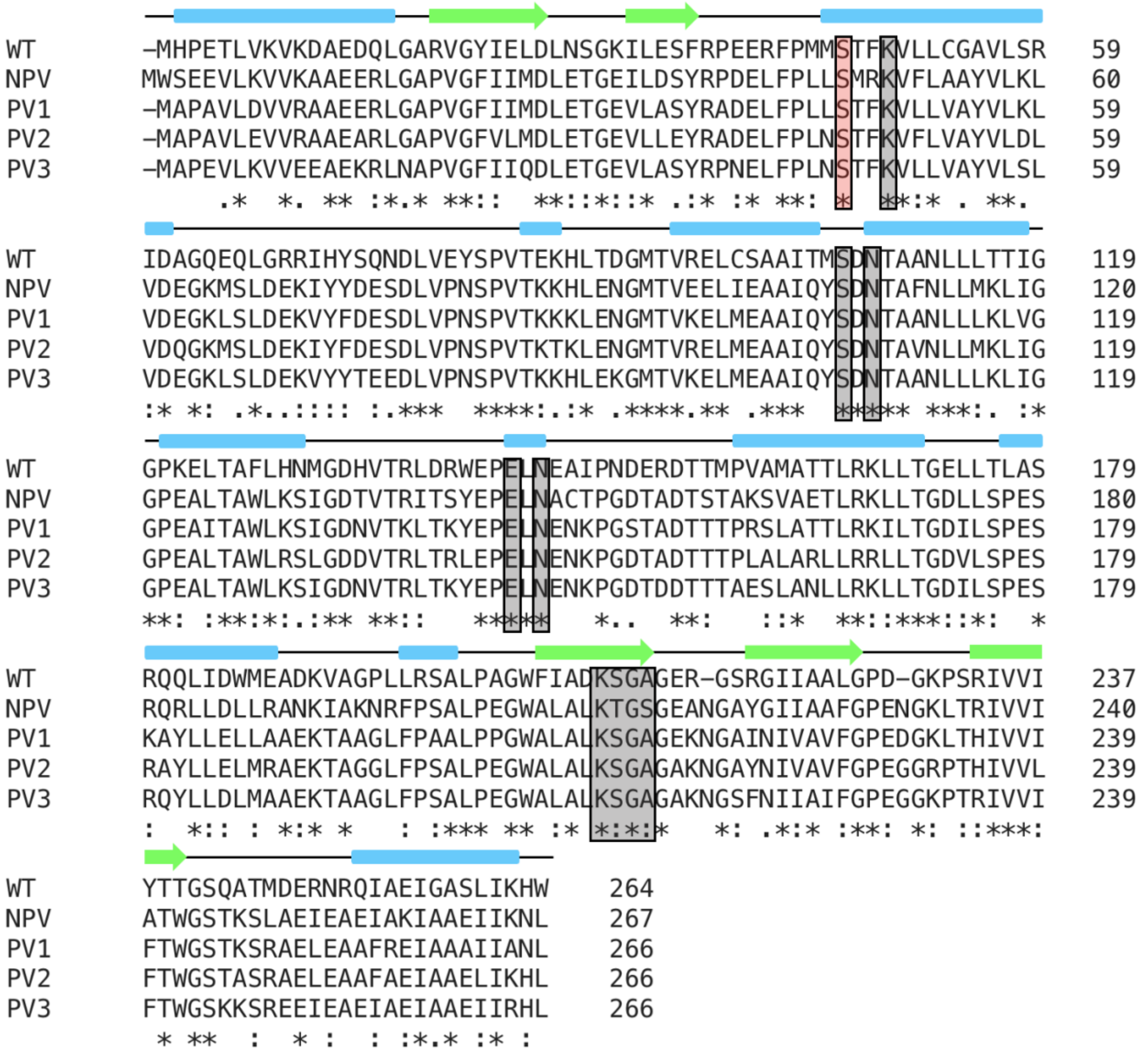
As a proof-of-principle, we sought to test NeuroFold's performance in generating variants of a well-characterized and clinically important enzyme. β-lactamase from *Escherichia coli* is a small, monomeric protein with an α-β fold, which catalyzes the hydrolysis of the β-lactam ring of β-lactam antibiotics such as penicillins and cephalosporins (Herzberg and Moult, 1987; Strynadka et al., 1992). This simplicity, in addition to the lack of cofactor required for structure and/or catalysis, makes β-lactamase an ideal candidate for variant production and testing. The active site of β-lactamase is composed of a set of conserved residues implicated in substrate binding and catalysis, including Ser70, Lys73, Ser130, Asn132, Glu166, Asn170, Lys234, Ser235, Gly236 and Ala237 (standard numbering scheme from Ambler et al. (1991) for class A β-lactamases). Ser70 is the nucleophile, directly participating in catalysis and forms the oxyanion hole with Ala237 to stabilize the negative charge formed by the tetrahedral intermediate during acylation and deacylation (Strynadka et al., 1992; Fisher and Mobashery, 2009). Lys73, Asn170, Glu166 and a water molecule are thought to activate Ser70 by abstracting a proton (Herzberg and Moult, 1987; Adachi et al., 1991; Delaire et al., 1991; Escobar et al., 1991; Strynadka et al., 1992; Damblon et al., 1996; Minasov et al., 2002; Meroueh et al., 2005). Ser130 shuttles protons between Lys73 and the leaving group nitrogen (Strynadka et al., 1992). Asn132, Lys234, Ser235 and Gly236 are involved in substrate binding and transition state stabilization (Strynadka et al., 1992; Delmas et al., 2010; Fonseca et al., 2012). Mutations at these residues have all been demonstrated to reduce activity (Palzkill, 2018).

ProteinMPNN was used to generate 4,096 enzyme variants of β-lactamase (PDB ID: 1BTL) using the Neurosnap platform (Neurosnap Inc., 2022; https://neurosnap.ai/). The resultant variants were then filtered using NeuroFold to produce a set of potential candidates. Of these potential candidates, four were randomly selected (see Table S3) for experimentation and their sequence alignment and predicted topology are presented in Figure 6. Three of the variants (i.e. Preserved Variants or PV) had their catalytic sites fixed during the ProteinMPNN sampling process. The catalytic sites were fixed in order to bias ProteinMPNN as well as to maximize the probability of obtaining functional enzymes. This step could be unnecessary as even samples without catalytic site preservation (i.e. Non-Preserved Variant or NPV) subsequently demonstrated partial conservation of the active site residues, with mutations similar to the variation seen across class A β-lactamases.

Remarkably, all variants selected for experimental validation exhibited sequence identities to the wildtype enzyme ranging from 49% to 56% (Figure 6) and mutations were evenly distributed throughout the entire protein sequence. Despite low sequence identity, the predicted secondary and tertiary structures of the variants were nearly identical (Figure 6 and Figure 7A). The amino acid composition, predicted pI (5.14 - 5.94) and predicted molecular weight (29.3 - 29.9 kDa) were also very similar (Table S4), thus demonstrating that the NeuroFold pipeline is effective at producing variants with low sequence identity, while conserving important physicochemical characteristics.

Biophysical Characterization of β-Lactamase Variant Structure and Function

The wild type (WT) β-lactamase and four candidate variants were recombinantly expressed in *E. coli* and purified using immobilized metal affinity chromatography (IMAC) from the soluble fraction. The variants were purified with comparable yields to the WT and migrated similarly when analyzed by SDS-PAGE, with molecular weights of approximately 30 kDa (Figure S2). Each variant had α-helical secondary structure comparable to the WT, as indicated by the far-UV circular dichroism (CD) spectra (Figure 7B-F) with characteristic minima at 208 and 222 nm, apparent maxima below 200 nm, and in agreement with structural predictions using AlphaFold2 (Figure 7A). Light scattering and flattening effects did not influence or distort the spectra within the range of wavelengths shown in Figure 7B-F. Preliminary stability data using far-UV CD



```
WT      -MHPETLVKVKDAEDQLGARVGYIELDLNSGKILESFRPEERFPMMSTFKVLLCGAVLSR   59
NPV     MWSEEVLKVVKAAEERLGAPVGFIIMDLETGEILDSYRPDELFPLLSMRKVFLAAYVLKL   60
PV1     -MAPAVLDVVRAAEERLGAPVGFIIMDLETGEVLASYRADELFPLLSTFKVLLVAYVLKL   59
PV2     -MAPAVLEVVRAAEARLGAPVGFVLMDLETGEVLLEYRADELFPLNSTFKVFLVAYVLDL   59
PV3     -MAPEVLKVVEEAEKRLNAPVGFIIQDLETGEVLASYRPNELFPLNSTFKVLLVAYVLSL   59
             .*  *. ** :*.* **::  **::*::* .:* :* **: *  **:* . **.

WT      IDAGQEQLGRRIHYSQNDLVEYSPVTEKHLTDGMTVRELCSAAITMSDNTAANLLLTTIG   119
NPV     VDEGKMSLDEKIYYDESDLVPNSPVTKKHLENGMTVEELIEAAIQYSDNTAFNLLMKLIG   120
PV1     VDEGKLSLDEKVYFDESDLVPNSPVTKKKLENGMTVKELMEAAIQYSDNTAANLLLKLVG   119
PV2     VDQGKMSLDEKIYFDESDLVPNSPVTKTKLENGMTVRELMEAAIQYSDNTAVNLLMKLIG   119
PV3     VDEGKLSLDEKVYYTEEDLVPNSPVTKKHLEKGMTVKELMEAAIQYSDNTAANLLLKLIG   119
        :* *: .*..:::: :.***  ****:.:* .****.** .***  ***** ***:. :*

WT      GPKELTAFLHNMGDHVTRLDRWEPELNEAIPNDERDTTMPVAMATTLRKLLTGELLTLAS   179
NPV     GPEALTAWLKSIGDTVTRITSYEPELNACTPGDTADTSTAKSVAETLRKLLTGDLLSPES   180
PV1     GPEAITAWLKSIGDNVTKLTKYEPELNENKPGSTADTTTPRSLATTLRKILTGDILSPES   179
PV2     GPEALTAWLRSLGDDVTRLTRLEPELNENKPGDTADTTTPLALARLLRRLLTGDVLSPES   179
PV3     GPEALTAWLKSIGDNVTRLTKYEPELNENKPGDTDDTTTAESLANLLRKLLTGDILSPES   179
        **: :**:*:.:** **::   *****   *..  **:   ::*  **::***::*:  *

WT      RQQLIDWMEADKVAGPLLRSALPAGWFIADKSGAGER-GSRGIIAALGPD-GKPSRIVVI   237
NPV     RQRLLDLLRANKIAKNRFPSALPEGWALALKTGSGEANGAYGIIAAFGPENGKLTRIVVI   240
PV1     KAYLLELLAAEKTAAGLFPAALPPGWALALKSGAGEKNGAINIVAVFGPEDGKLTHIVVI   239
PV2     RAYLLELMRAEKTAGGLFPSALPEGWALALKSGAGAKNGAYNIVAVFGPEGGRPTHIVVL   239
PV3     RQYLLDLMAAEKTAAGLFPSALPEGWALALKSGAGAKNGSFNIIAIFGPEGGKPTRIVVI   239
        :  *:: : *:* *   : :*** ** :* *:*:*   *: .*:* :**: *: ::***:

WT      YTTGSQATMDERNRQIAEIGASLIKHW    264
NPV     ATWGSTKSLAEIEAEIAKIAAEIIKNL    267
PV1     FTWGSTKSRAELEAAFREIAAAIIANL    266
PV2     FTWGSTASRAELEAAFAEIAAELIKHL    266
PV3     FTWGSKKSREEIEAEIAEIAAEIIRHL    266
         * **  :  * :  : :*.* :* :
```

Figure 6. Amino acid sequence alignment of β-lactamase and variants designed by NeuroFold. CLUSTAL multiple sequence alignment by MUSCLE (3.8) using the EMBL-EBI platform (Edgar, 2004). An * (asterisk) indicates positions which have a single, fully conserved residue. A : (colon) indicates conservation between groups of residues with strongly similar properties. A . (period) indicates conservation between groups of residues with weakly similar properties. Secondary structure is mapped based on the crystal structure of WT β-lactamase (PDB: 1BTL) and AlphaFold2 predictions for β-lactamase variants (NPV - Non-Preserved Variant; PV - Preserved Variant). Blue blocks represent α-helices and green arrows represent β-strands. Gray highlights correspond to active site residues that were optionally preserved; the catalytic serine residue is highlighted in red.

spectroscopy above the melting point of WT β-lactamase (70℃) suggests that the variants are more thermostable than the WT, indicated by a reduction in the ellipticity at 222 nm for the WT relative to the variants (Figure 7B-F). This data supports that NeuroFold is capable of generating variants that maintain the structure and increase the stability of the WT protein backbone.

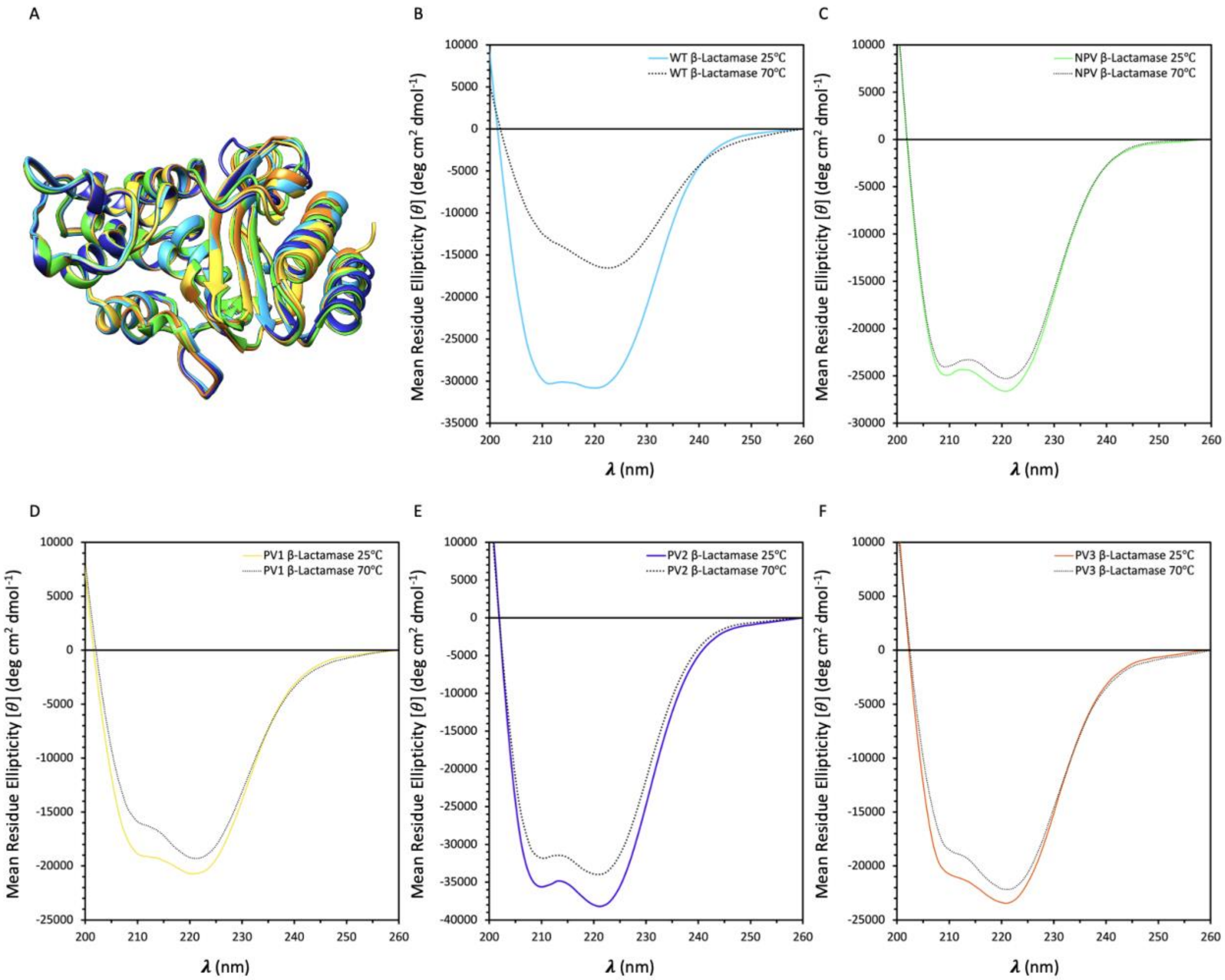


Figure 7. Biophysical Characterization of WT and Variant β-Lactamase Structure and Stability. (A) Structural alignment of WT β-Lactamase (PDB #1BTL) and AlphaFold2 predicted structures for β-Lactamase variants using Chimera MatchMaker, which are in close agreement at the backbone level. (B-F) Far-ultraviolet circular dichroism (CD) spectra at 25℃ (solid and coloured) and 70℃ (dotted and black) for WT (B), non-preserved variant (NPV) (C), preserved variant 1 (PV1) (D), preserved variant 2 (PV2) (E) and preserved variant 3 (PV3) (F) β-Lactamases. All spectra indicate close agreement of secondary structure (α-helical), where WT β-Lactamase structure is destabilized at 70℃, marked by a large reduction in the relative ellipticity at 222 nm compared to the variant β-Lactamases.

The β-lactamase activity assay was used to assess the enzymatic activity of the variants relative to the wild type (Figure 8A and B). The WT exhibited a specific activity of 6.28 a.u. min^{-1} mg^{-1} at 413 nm for the substrate penicillin G. The non-preserved variant (NPV) exhibited a specific activity of 4.58 a.u. min^{-1} mg^{-1} at 413 nm for the substrate penicillin G. The preserved variants (PV1-3) exhibited specific activities of 0.628, 2.56 and 0.048 a.u. min^{-1} mg^{-1} at 413 nm for the substrate penicillin G, respectively. The NPV, which possessed mutations to some active site residues (A212S and S210T), demonstrated 73% relative specific activity to the WT. PV2 exhibited 41% relative specific activity to the WT, whereas PV1 and PV3 exhibited relative specific activities below 10% compared to the WT. This unpredictability remains one of the major limitations of rational protein design, as it is often the case that mutations far from the active site of an enzyme can have measurable effects on activity (Stiffler et al., 2015). Key catalytic residues in the active site of WT β-lactamase reported in the literature (Palzkill, 2018) were aligned to compare their orientation (Figure 8C-G). The active sites showed very close agreement at the backbone and residue level, even for the NPV. However, minor differences in binding pocket shape, electrostatic surface and substrate binding could be observed between the variants and relative to the WT (Figure 8H-L). These changes suggest that, although the active site residues remain unchanged, substrate binding or transition state stabilization could vary considerably between the variants and relative to the WT, which could explain the significant differences in activity. The results illustrate an important trade-off in the practice of enzyme design – changes to sequence which might confer enhanced stability or other favorable properties often come at the cost of decreased catalytic efficiency. Nonetheless, we show that NeuroFold has the ability to generate enzyme variants that maintain varying levels of activity with less than 56% sequence identity.

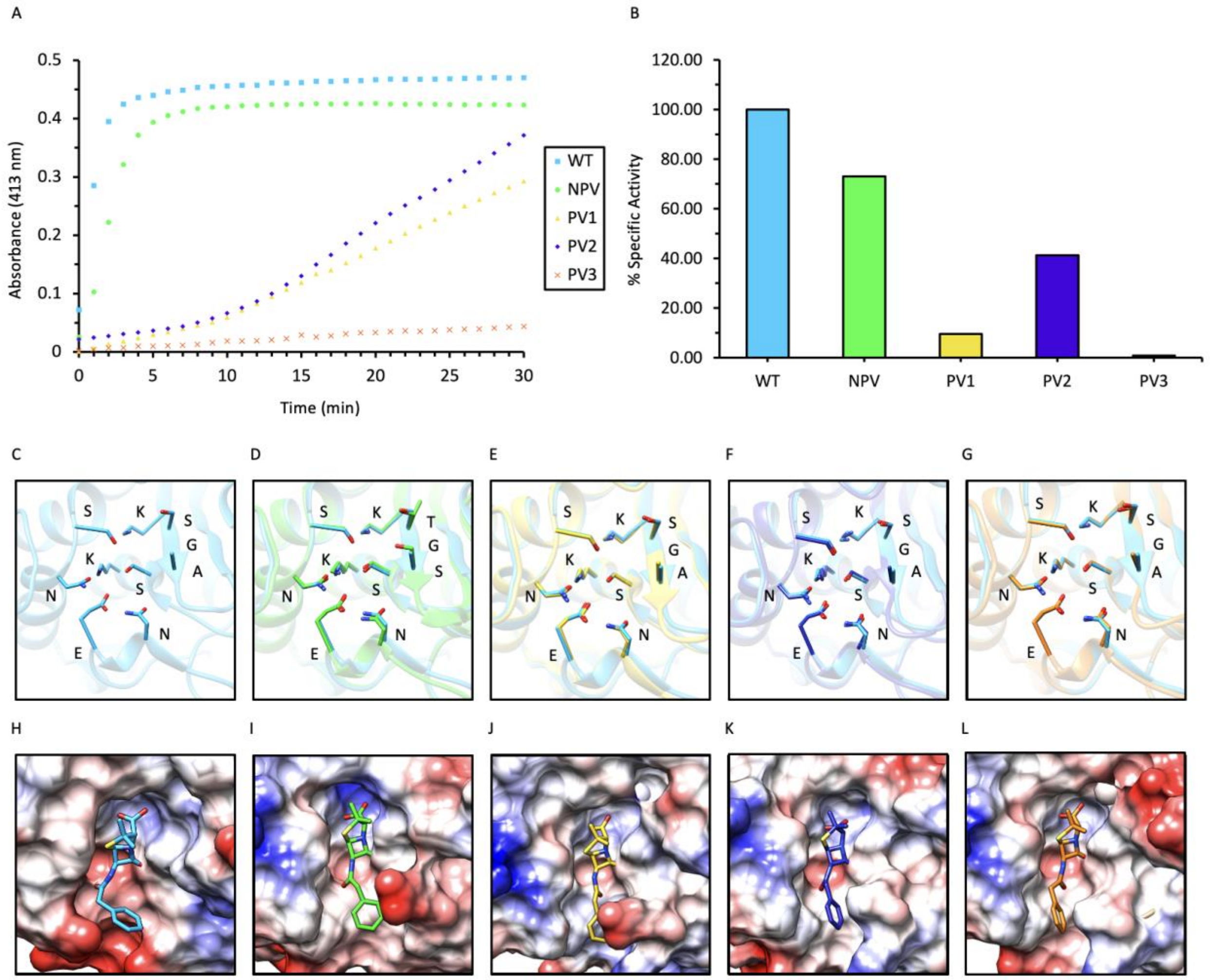


Figure 8. Enzymatic Activity and Computational Characterization of WT and Variant β-Lactamases. (A) Enzyme activity of WT and variant β-lactamases measured as change in absorbance at 413 nm against time (minutes) in the presence of penicillin G substrate and the PAR-2Hg^{2+} complex. An increase in absorbance at 413 nm is a result of the hydrolysis of penicillin G to penicilloic acid, which chelates Hg^{2+} from the PAR-2Hg^{2+} complex, causing a color change from red to yellow. (B) % Specific Activity of variant β-Lactamases relative to the WT determined by linear regression of the tangent line to each curve in (A) at

time = 0 min. (C-G) Structural alignment of key active site residues between the WT (C), non-preserved variant (NPV) (D), preserved variant 1 (PV1) (E), preserved variant 2 (PV2) (F) and preserved variant 3 (PV3) (G) β-Lactamases which are in close agreement at the backbone and residue level. (H-L) Binding pocket shape, electrostatic surface and substrate (penicillin G) docking comparisons between the WT (H), NPV (I), PV1 (J), PV2 (K) and PV3 (L) β-Lactamases which indicate minor differences in binding pocket shape, electrostatic surface and substrate binding.

Concluding Remarks

In this study, we describe NeuroFold, a multimodal model that is able to integrate sequence, structural and evolutionary information to make inferences on the global protein space and to design functional enzyme variants. In particular, we show that NeuroFold can be used to generate divergent versions of the enzyme β-lactamase that exhibit enhanced thermostability and measurable activity levels up to 73% relative to the WT. Despite the remarkable capabilities of NeuroFold, there are certain limitations to consider. Firstly, it should be noted that NeuroFold may encounter challenges when dealing with certain enzyme groups (e.g. orphaned proteins, multimeric proteins, very large proteins) that exhibit incompatible formats not yet supported by the model. Moreover, NeuroFold's current limitations prevent it from effectively producing variants of large enzymes greater than 2,500 amino acids in length, primarily due to the limitations in VRAM and compute time requirements. Finally, NeuroFold has not yet been sufficiently tested on proteins that contain metal-binding sites, allosteric binding domains, or multimeric enzymes. These limitations highlight areas where further development and improvement of NeuroFold are required to expand its applicability across a broader range of enzymes and molecular configurations.

Materials and Methods

β-Lactamase Variant Structure Prediction and Docking with Penicillin G

The structures of each enzyme variant and the wildtype were predicted with AlphaFold2's ColabFold implementation using the Neurosnap platform (Jumper et al., 2021; Mirdita et al., 2022; Neurosnap Inc., 2022; https://neurosnap.ai/). Molecular docking experiments against penicillin G were conducted using the DiffDock model on the Neurosnap platform (Corso et al., 2023; Neurosnap Inc., 2022; https://neurosnap.ai/). See Table 2 for details.

Table 2. Predicted structures and molecular docking with the target substrate. Structure prediction of the enzyme variants and 1BTL are performed using AlphaFold2 on the Neurosnap platform. DiffDock, a state of the art molecular docking model is used to validate whether the enzyme substrate, penicillin G, interacts with the enzyme's active site. Each prediction was performed on the Neurosnap platform and thus all the results, inputs, and configurations are available at the job share links.

Job Description	Neurosnap Job Results Page
AlphaFold2 predictions of WT β-lactamase and its variants.	https://neurosnap.ai/job/655c1f02c43c0cfb6fdfd577?share=656651f702c4834793f4c2b7
DiffDock WT and penicillin G.	https://neurosnap.ai/job/656556b602c4834793f4c198?share=6567a82b02c4834793f4c3b8
DiffDock PV1 and penicillin G.	https://neurosnap.ai/job/656556f602c4834793f4c19b?share=6567a82e02c4834793f4c3b9
DiffDock PV2 and penicillin G.	https://neurosnap.ai/job/6565571a02c4834793f4c19e?share=6567a83302c4834793f4c3ba
DiffDock PV3 and penicillin G.	https://neurosnap.ai/job/656559d402c4834793f4c1a1?share=6567a83502c4834793f4c3bb
DiffDock NPV and penicillin G.	https://neurosnap.ai/job/656559ea02c4834793f4c1a4?share=656651ed02c4834793f4c2b6

Gene Construction

Synthetic cDNA sequences encoding for the wild-type *E. coli* β-lactamase (PDB 1BTL) and 4 variants designed using NeuroFold were synthesized and cloned into pET28(a)+ vectors by Twist Bioscience (https://www.twistbioscience.com/). The endogenous N-terminal transit peptide was removed from the wild-type sequence before variants were generated, so this was not present in any of the constructs. Constructs were designed to include a C-terminal poly-histidine tag fusion for affinity purification and the pET28(a)+ vector provided a kanamycin resistance gene for selection of bacterial clones.

Protein Expression and Purification

Constructs were introduced into *E. coli* (DH5α) and positive clones were isolated, plasmid DNA purified, and plasmids were transformed into *E. coli* BL21 (DE3) LOBSTR cells. Overnight cultures were prepared in LB broth containing Kanamycin from a single colony and incubated at 37°C overnight while shaking at 250 rpm. Overnight cultures were used to inoculate a 1 L culture of LB+Kan and incubated at 37°C for 3-5 hours, until an OD_{600} of 0.6-0.8 was reached. Expression was induced with 0.1 mM IPTG for 5 hours at 37°C. Cultures were centrifuged for 15 minutes at 4°C and 8,000 x g to pellet the bacterial cells. The supernatant was discarded, and the cells resuspended in lysis buffer (20 mM Tris-HC, pH 8.0, 150 mM NaCl, 5 mM $MgCl_2$, 15 mM imidazole, 1X cOmplete protease inhibitor cocktail (Roche). Cells were incubated in lysis buffer for 1 hour before physical lysis using a Constant Cell Disruptor operating at 20 kPSi. The lysate was centrifuged for 20 minutes at 4°C and 20,000 x g to pellet inclusion bodies and unlysed cells. The supernatant containing the clarified lysate was applied to a pre-equilibrated (with 10 column volumes of lysis buffer) immobilized metal affinity chromatography (IMAC) gravity flow column using Ni-NTA HisPur resin from Bio-Rad with a 1 mL column volume. The clarified lysate was incubated with the resin for 30 minutes at 4°C before collecting the flow-through. The column was washed with 10 column volumes of wash buffer (20 mM Tris-HCl, pH 8.0, 150 mM NaCl, 30 mM imidazole) and the protein was eluted with 10 column volumes of elution buffer (20 mM Tris-HCl, pH 8.0, 150 mM NaCl, 250 mM imidazole) in 1 mL increments. Eluted proteins were desalted to remove imidazole and reduce NaCl to 50 mM (20 mM Tris-HCl, pH 8.0, 50 mM NaCl) using Econo-Pac 10DG Desalting Columns (Bio-Rad), and all fractions were analyzed by Bradford protein assay and SDS-PAGE (4% stacking and 12% resolving gels, stained by Coomassie Brilliant Blue).

β-Lactamase Activity Assay

A colorimetric assay for β-lactamase activity using penicillin G and the PAR-$2Hg^{2+}$ complex was performed as described by Lee et al. (2017). A solution containing PAR (20 μM), Hg^{2+} (40 μM) and penicillin G (100 μM) in Tris-HCl buffer (20 mM, pH 8.0) was incubated for 40 minutes. After the addition of protein (to a final concentration of 3 μM), the absorbance of the solution was recorded in 1 minute increments for 30 minutes at 413 nm using a BioTek Synergy HT spectrophotometer to determine the change in absorbance/minute per mg of protein.

Circular Dichroism (CD) Spectroscopy

Far-UV CD spectra of the WT β-lactamase and the 4 variants (3-5 μM) in CD buffer (20 mM Tris-HCl, pH 8.0, 50 mM NaCl) were measured at 25 ºC from 190-260 nm using an AVIV 215 spectropolarimeter. Measurements were carried out in a 0.1 cm path-length quartz cuvette at 1 nm resolution. The reported spectra are an average of 9 scans. CD experiments were repeated at 70℃ to monitor structural stability.

Data & Code Availability Statement

All computed structures, MSAs, and model metrics are freely available for download using the following URL https://neurosnap.ai/neurofold/dataset. Additionally, data can be made available upon reasonable request and by providing the authors with a complimentary lunch.

Author Contributions

K.A. developed (coded and trained) NeuroFold; conceptualized and performed *in silico* experiments including variant generation, characterization and docking experiments. M.F. conceptualized and performed *in vitro* experiments including wild-type protein selection, construct design, protein expression and purification, activity assays and CD analysis of protein structure and stability. K.A. and M.F. wrote corresponding sections of the manuscript. M.D.S. and C.D.M.C. participated in review and editing of the manuscript in addition to intellectual contributions, funding, and supervision throughout the study. All authors have read and agree to the submitted version of the manuscript.

Acknowledgements

This research was funded by Natural Sciences and Engineering Research Council of Canada (NSERC); postgraduate doctoral scholarship awarded to M.F. (PGSD3-2021-559542) and Discovery Grants awarded to M.D.S. (RGPIN-2017-05437) and C.D.M.C. (RGPIN-2020-04117) and a Discovery Development Grant awarded to M.D.S. (DDG-2023-00034).

Ethics Declarations

K.A. is the CEO & Founder of Neurosnap Inc. (https://neurosnap.ai/) and C.D.M.C. sits on the advisory board as a scientific advisor.

Conflicts of Interest

The funders had no role in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish results. K.A. is the C.E.O. and Founder of Neurosnap Inc. and C.D.M.C. sits on the advisory board as a scientific advisor. Neurosnap Inc. will benefit financially from the development and validation of NeuroFold as a tool in its suite of bioinformatic services.

References

Ambler, R. P., Coulson, A., Frère, J., Ghuysen, J., Joris, B., Forsman, M., Lévesque, R. C., Tiraby, G., & Waley, S. G. (1991). A standard numbering scheme for the class A β-lactamases. *Biochemical Journal*, *276*(1), 269–270. https://doi.org/10.1042/bj2760269

Baek, M., & Baker, D. (2022). Deep learning and protein structure modeling. *Nature Methods*, *19*(1), 13–14. https://doi.org/10.1038/s41592-021-01360-8

Baek, M., DiMaio, F., Anishchenko, I., Dauparas, J., Овчинников, С. Г., Lee, G. R., Wang, J., Cong, Q., Kinch, L. N., Schaeffer, R. D., Millán, C., Park, H., Adams, C., Glassman, C. R., DeGiovanni, A., Pereira, J., Rodrigues, A. V., Van Dijk, A. A., Ebrecht, A. C., . . . Baker, D. (2021). Accurate prediction of protein structures and interactions using a three-track neural network. *Science*, *373*(6557), 871–876. https://doi.org/10.1126/science.abj8754

Chan, Y., Venev, S. V., Zeldovich, K. B., & Matthews, C. R. (2017). Correlation of fitness landscapes from three orthologous TIM barrels originates from sequence and structure constraints. *Nature Communications*, *8*(1). https://doi.org/10.1038/ncomms14614

Chiasson, M. A., Rollins, N., Stephany, J. J., Sitko, K. A., Matreyek, K. A., Verby, M., Sun, S., Roth, F. P., DeSloover, D., Marks, D., Rettie, A. E., & Fowler, D. M. (2020). Multiplexed measurement of variant abundance and activity reveals VKOR topology, active site and human variant impact. *eLife*, *9*. https://doi.org/10.7554/elife.58026

Corso, G., Stark, H., Jing, B., Barzilay, R., & Jaakkola, T. S. (2023). DiffDock: diffusion steps, twists, and turns for molecular docking. *Zenodo (CERN European Organization for Nuclear Research)*. https://doi.org/10.48550/arxiv.2210.01776

Damblon, C., Raquet, X., Lian, L., Lamotte-Brasseur, J., Fonzé, E., Charlier, P., Roberts, G. C. K., & Frère, J. (1996). The catalytic mechanism of beta-lactamases: NMR titration of an active-site lysine residue of the TEM-1 enzyme. *Proceedings of the National Academy of Sciences of the United States of America*, *93*(5), 1747–1752. https://doi.org/10.1073/pnas.93.5.1747

Dauparas, J., Anishchenko, I., Bennett, N. R., Bai, H., Ragotte, R. J., Milles, L. F., Wicky, B. I. M., Courbet, A., De Haas, R. J., Bethel, N., Leung, P. J. Y., Huddy, T. F., Pellock, S., Tischer, D., Chan, F., Koepnick, B., Nguyen, H., Kang, A., Sankaran, B., . . . Baker, D. (2022). Robust deep learning–based protein sequence design using ProteinMPNN. *Science*, *378*(6615), 49–56. https://doi.org/10.1126/science.add2187

Delaire, M., Lenfant, F., Labia, R., & Masson, J. (1991). Site-directed mutagenesis on TEM-1 ß-lactamase: role of Glul66 in catalysis and substrate binding. *Protein Engineering Design & Selection*, *4*(7), 805–810. https://doi.org/10.1093/protein/4.7.805

Delmas, J., Leyssene, D., Dubois, D., Birck, C., Vazeille, E., Robin, F., & Bonnet, R. (2010). Structural Insights into Substrate Recognition and Product Expulsion in CTX-M Enzymes. *Journal of Molecular Biology*, *400*(1), 108–120. https://doi.org/10.1016/j.jmb.2010.04.062

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, *32*(5), 1792–1797. https://doi.org/10.1093/nar/gkh340

Escobar, W. A., Tan, A. K., & Fink, A. L. (1991). Site-directed mutagenesis of .beta.-lactamase leading to accumulation of a catalytic intermediate. *Biochemistry*, *30*(44), 10783–10787. https://doi.org/10.1021/bi00108a025

Fonseca, F., Chudyk, E. I., Van Der Kamp, M. W., Correia, A., Mulholland, A. J., & Spencer, J. (2012). The Basis for Carbapenem Hydrolysis by Class A β-Lactamases: A Combined Investigation using Crystallography and Simulations. *Journal of the American Chemical Society*, *134*(44), 18275–18285. https://doi.org/10.1021/ja304460j

Herzberg, O., & Moult, J. (1987). Bacterial Resistance to β-Lactam Antibiotics: Crystal Structure of β-Lactamase from Staphylococcus aureus PC1 at 2.5 Å Resolution. *Science*, *236*(4802), 694–701. https://doi.org/10.1126/science.3107125

Hsu, C. C., Verkuil, R., Liu, J., Lin, Z., Hie, B., Sercu, T., Lerer, A., & Rives, A. (2022). Learning inverse folding from millions of predicted structures. *bioRxiv (Cold Spring Harbor Laboratory)*. https://doi.org/10.1101/2022.04.10.487779

Johnson, S. R., Fu, X., Viknander, S., Goldin, C., Monaco, S., Zelezniak, A., & Yang, K. (2023). Computational scoring and experimental evaluation of enzymes generated by neural networks. *bioRxiv (Cold Spring Harbor Laboratory)*. https://doi.org/10.1101/2023.03.04.531015

Johnson, S. R., Massie, K., Monaco, S., & Sayed, Z. (2021). Generating novel protein sequences using Gibbs sampling of masked language models. *bioRxiv (Cold Spring Harbor Laboratory)*. https://doi.org/10.1101/2021.01.26.428322

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. a. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., . . . Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, *596*(7873), 583–589. https://doi.org/10.1038/s41586-021-03819-2

Koehl, P., & Levitt, M. (2002). Sequence Variations within Protein Families are Linearly Related to Structural Variations. *Journal of Molecular Biology*, *323*(3), 551–562. https://doi.org/10.1016/s0022-2836(02)00971-3

Li, C., Zhang, R., Wang, J., Wilson, L. M., & Yan, Y. (2020). Protein engineering for improving and diversifying natural product biosynthesis. *Trends in Biotechnology*, *38*(7), 729–744. https://doi.org/10.1016/j.tibtech.2019.12.008

Llarrull, L. I., Fisher, J. F., & Mobashery, S. (2009). Molecular Basis and Phenotype of Methicillin Resistance in Staphylococcus aureus and Insights into New β-Lactams That Meet the Challenge. *Antimicrobial Agents and Chemotherapy*, *53*(10), 4051–4063. https://doi.org/10.1128/aac.00084-09

Madani, A., Krause, B., Greene, E. R., Subramanian, S., Mohr, B. P., Holton, J. M., Olmos, J., Xiong, C., Sun, Z. Z., Socher, R., Fraser, J. S., & Naik, N. (2023). Large language models generate functional protein sequences across diverse families. *Nature Biotechnology*, *41*(8), 1099–1106. https://doi.org/10.1038/s41587-022-01618-2

Matreyek, K. A., Starita, L. M., Stephany, J. J., Martin, B., Chiasson, M. A., Gray, V. E., Kircher, M., Khechaduri, A., Dines, J. N., Hause, R. J., Bhatia, S., Evans, W. E., Relling, M. V., Yang, W., Shendure, J., & Fowler, D. M. (2018). Multiplex assessment of protein variant abundance by massively parallel sequencing. *Nature Genetics*, *50*(6), 874–882. https://doi.org/10.1038/s41588-018-0122-z

Meier, J., Rao, R., Verkuil, R., Liu, J., Sercu, T., & Rives, A. (2021). Language models enable zero-shot prediction of the effects of mutations on protein function. *bioRxiv (Cold Spring Harbor Laboratory)*. https://doi.org/10.1101/2021.07.09.450648

Meroueh, S. O., Fisher, J. F., Schlegel, H. B., & Mobashery, S. (2005). Ab initio QM/MM study of Class A Β-Lactamase acylation: dual participation of GLU166 and LYS73 in a concerted base promotion of SER70. *Journal of the American Chemical Society*, *127*(44), 15397–15407. https://doi.org/10.1021/ja051592u

Minasov, G., Wang, X., & Shoichet, B. K. (2002). An ultrahigh resolution structure of TEM-1 Β-Lactamase suggests a role for GLU166 as the general base in acylation. *Journal of the American Chemical Society*, *124*(19), 5333–5340. https://doi.org/10.1021/ja0259640

Mirdita, M., Schütze, K., Moriwaki, Y., Heo, L., Овчинников, С. Г., & Steinegger, M. (2022). ColabFold: making protein folding accessible to all. *Nature Methods*, *19*(6), 679–682. https://doi.org/10.1038/s41592-022-01488-1

Neurosnap Inc. (2022). *Neurosnap - Computational Biology Platform for Research.* neurosnap.ai. https://neurosnap.ai/

Notin, P., Dias, M., Frazer, J., Marchena-Hurtado, J., Gomez, A. N., Marks, D. S., & Gal, Y. (2022). Tranception: protein fitness prediction with autoregressive transformers and inference-time retrieval. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2205.13760

Palzkill, T. (2018). Structural and Mechanistic Basis for Extended-Spectrum Drug-Resistance Mutations in Altering the Specificity of TEM, CTX-M, and KPC β-lactamases. *Frontiers in Molecular Biosciences*, *5*. https://doi.org/10.3389/fmolb.2018.00016

Plotly Technologies Inc. (2015). *Collaborative Data science*. https://plot.ly/

Rao, R., Bhattacharya, N., Thomas, N., Duan, Y., Chen, X., Canny, J., Abbeel, P., & S. Song, Y. (2019). Evaluating Protein Transfer Learning with TAPE. *arXiv*. https://doi.org/10.48550/arXiv.1906.08230

Rao, R., Liu, J., Verkuil, R., Meier, J., Canny, J., Abbeel, P., Sercu, T., & Rives, A. (2021). MSA Transformer. *bioRxiv (Cold Spring Harbor Laboratory)*. https://doi.org/10.1101/2021.02.12.430858

Repecka, D., Jauniškis, V., Karpus, L., Rembeza, E., Rokaitis, I., Zrimec, J., Povilonienė, S., Laurynėnas, A., Viknander, S., Abuajwa, W., Savolainen, O., Meškys, R., Engqvist, M. K. M., & Zelezniak, A. (2021). Expanding functional protein sequence spaces using generative adversarial networks. *Nature Machine Intelligence*, *3*(4), 324–333. https://doi.org/10.1038/s42256-021-00310-5

Robinson, P. (2015). Enzymes: principles and biotechnological applications. *Essays in Biochemistry*, *59*, 1–41. https://doi.org/10.1042/bse0590001

Rocklin, G. J., Chidyausiku, T. M., Goreshnik, I., Ford, A. T., Houliston, S., Lemak, A., Carter, L., Ravichandran, R., Mulligan, V. K., Chevalier, A., Arrowsmith, C., & Baker, D. (2017). Global analysis of protein folding using massively parallel design, synthesis, and testing. *Science*, *357*(6347), 168–175. https://doi.org/10.1126/science.aan0693

Romero, P. A., & Arnold, F. H. (2009). Exploring protein fitness landscapes by directed evolution. *Nature Reviews Molecular Cell Biology*, *10*(12), 866–876. https://doi.org/10.1038/nrm2805

Russ, W. P., Figliuzzi, M., Stöcker, C., Barrat-Charlaix, P., Socolich, M., Kast, P., Hilvert, D., Monasson, R., Cocco, S., Weigt, M., & Ranganathan, R. (2020). An evolution-based model for designing chorismate mutase enzymes. *Science*, *369*(6502), 440–445. https://doi.org/10.1126/science.aba3304

Saitou, N., & Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*. https://doi.org/10.1093/oxfordjournals.molbev.a040454

Site-directed mutants, at position 166, of RTEM-1 beta-lactamase that form a stable acyl-enzyme intermediate with penicillin. (1991, February 15). PubMed. https://pubmed.ncbi.nlm.nih.gov/1993691/

Steinegger, M. (2017). MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature Biotechnology*, *35*(11), 1026–1028. https://doi.org/10.1038/nbt.3988

Stiffler, M. A., Hekstra, D. R., & Ranganathan, R. (2015). Evolvability as a function of purifying selection in TEM-1 Β-Lactamase. *Cell*, *160*(5), 882–892. https://doi.org/10.1016/j.cell.2015.01.035

Strynadka, N., Adachi, H., Jensen, S. E., Johns, K., Sielecki, A. R., Betzel, C., Sutoh, K., & James, M. N. (1992). Molecular structure of the acyl-enzyme intermediate in β-lactam hydrolysis at 1.7 Å resolution. *Nature*, *359*(6397), 700–705. https://doi.org/10.1038/359700a0

Süzek, B. E., Huang, H., McGarvey, P. B., Mazumder, R., & Wu, C. H. (2007). UniRef: comprehensive and non-redundant UniProt reference clusters. *Bioinformatics*, *23*(10), 1282–1288. https://doi.org/10.1093/bioinformatics/btm098

Yang, K., Lu, A. X., & Fusi, N. (2022). Convolutions are competitive with transformers for protein sequence pretraining. *bioRxiv (Cold Spring Harbor Laboratory)*. https://doi.org/10.1101/2022.05.19.492714

Yang, K., Zanichelli, N., & Yeh, H. (2022). Masked Inverse Folding with Sequence Transfer for Protein Representation Learning. *bioRxiv (Cold Spring Harbor Laboratory)*. https://doi.org/10.1101/2022.05.25.493516

Supplementary Information

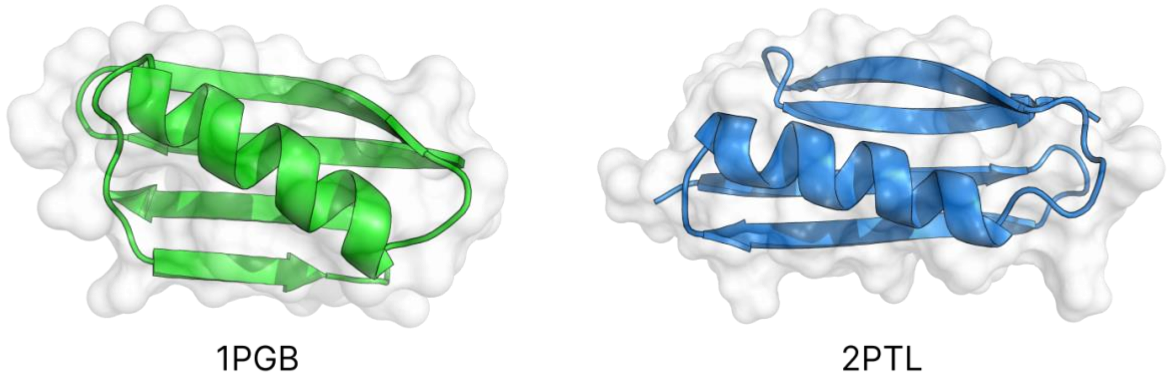


Figure S1. Comparison of 1PGB and 2PTL. The B1 immunoglobulin-binding domain of *streptococcal* protein (1PGB) and the B1 immunoglobulin light chain binding domain of *Peptostreptococcus magnus* (2PTL) have very similar folds despite a pairwise sequence identity of 14% (Koehl & Levitt, 2002).

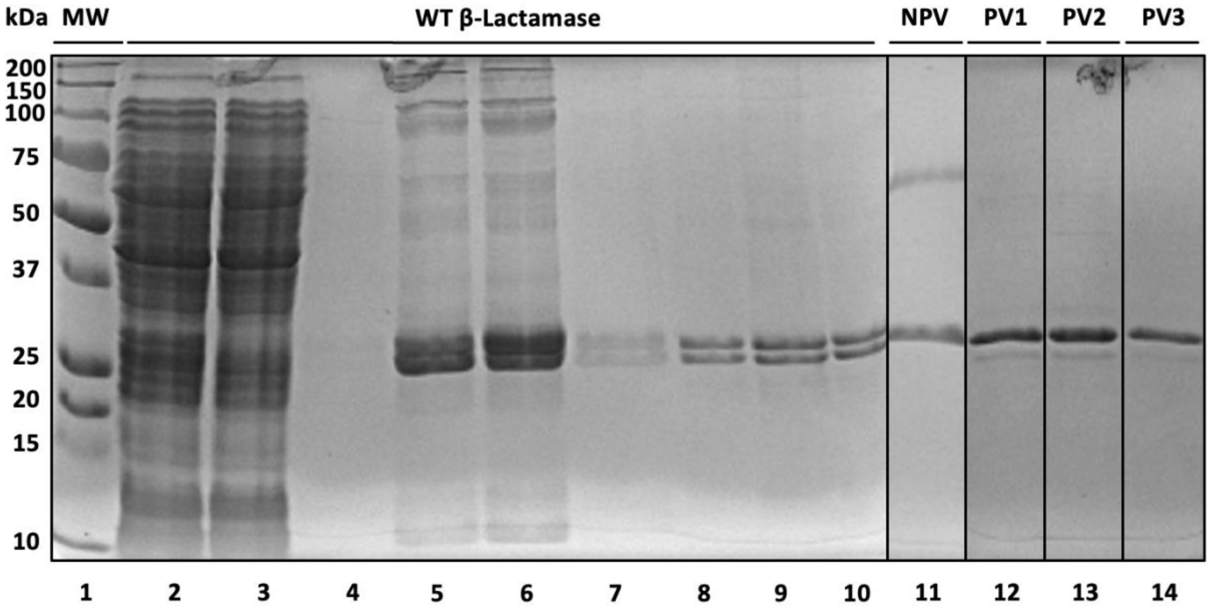


Figure S2. SDS-PAGE Analysis of Expression and Purification of β-Lactamase WT and Generated Variants. (Lane 1) Molecular weight markers in kDa. (Lane 2) *E. coli* whole cell lysate after inducing expression of the WT β-Lactamase. (Lane 3) Flow-Through after applying immobilized metal affinity chromatography (IMAC) with Ni-NTA resin and 5 mM imidazole. (Lane 4) Wash Fraction with 30 mM imidazole. (Lanes 5-7) Elution Fractions with 250 mM imidazole. (Lanes 8-10) Purified WT β-Lactamase after size exclusion chromatography. (Lanes 11-14) Purified β-Lactamase variants non-preserved (NPV), 1, (PV1), 2 (PV2) and 3 (PV3) after size exclusion chromatography. Purified proteins appear at the expected molecular weight (~30 kDa) between the 25 and 37 kDa markers. Proteins are separated by a 12% resolving gel and visualized by staining with Coomassie Brilliant Blue.

Table S1. ProteinMPNN Settings on Neurosnap

Setting	**Preserved Catalytic Sites**	**Non-Preserved Catalytic Sites**
PDB	1BTL	1BTL
Chains	A	A
Model Version	original	original
Model Type	v_48_020	v_48_020
Fixed Positions	70-74, 130-134, 166-171, 234-238	None
# Sequences	2048	2048
Sampling Temperature	0.01	0.01

Table S2. Calculating Average Success Rate of DMS Experiments

We use the ProteinGym dataset (Notin et al., 2022) to calculate the average success rate of a DMS experiment by getting the top 65th percentile of the subset of positives for each dataset to use as the reference sequence. All sequences with a DMS_score greater than the DMS_score of the reference sequence are considered passable. The total number of passing sequences divided by the total number of sequences within that DMS experiment is considered the success rate of that experiment. We calculate the average of the success rates and present them below alongside the average. The python code utilized to produce these numbers is also provided.

Dataset Name	**Success Rate**
Average of all Datasets	14.17%
KCNH2_HUMAN_Kozek_2020	17.50%
SCN5A_HUMAN_Glazer_2019	17.41%
NRAM_I33A0_Jiang_standard_2016	17.45%
ENV_HV1B9_DuenasDecamp_2016	9.87%
TPOR_HUMAN_Bridgford_S505N_2020	7.83%
VKOR1_HUMAN_Chiasson_activity_2020	22.24%
UBE4B_MOUSE_Starita_2013	29.25%
A0A1I9GEU1_NEIME_Kennouche_2019	17.46%
BLAT_ECOLX_Jacquier_2013	5.46%
P53_HUMAN_Kotler_2018	20.71%
CCDB_ECOLI_Adkar_2012	16.41%
GAL4_YEAST_Kitzman_2015	24.27%
RL401_YEAST_Roscoe_2013	22.18%
TADBP_HUMAN_Bolognesi_2019	17.47%
RL401_YEAST_Mavor_2016	24.02%
IF1_ECOLI_Kelsic_2016	24.29%
RL401_YEAST_Roscoe_2014	26.01%
P84126_THETH_Chan_2017	15.67%
TRPC_SACS2_Chan_2017	13.82%
TRPC_THEMA_Chan_2017	9.81%

DLG4_RAT_McLaughlin_2012	27.09%
TAT_HV1BR_Fernandes_2016	23.84%
POLG_HCVJF_Qi_2014	15.40%
CCDB_ECOLI_Tripathi_2016	0.00%
SUMO1_HUMAN_Weile_2017	21.59%
MTH3_HAEAE_Rockah-Shmuel_2015	25.38%
AACC1_PSEAI_Dandage_2018	17.49%
CALM1_HUMAN_Weile_2017	17.48%
PA_I34A1_Wu_2015	17.47%
BRCA1_HUMAN_Findlay_2018	25.97%
REV_HV1H2_Fernandes_2016	17.33%
ESTA_BACSU_Nutschel_2020	17.50%
A0A2Z5U3Z0_9INFA_Wu_2014	17.49%
DYR_ECOLI_Thompson_plusLon_2019	24.29%
SYUA_HUMAN_Newberry_2020	27.03%
UBC9_HUMAN_Weile_2017	17.48%
GCN4_YEAST_Staller_induction_2018	17.48%
VKOR1_HUMAN_Chiasson_abundance_2020	17.48%
NUD15_HUMAN_Suiter_2020	23.42%
Q59976_STRSQ_Romero_2015	22.97%
RASH_HUMAN_Bandaru_2017	24.31%
TPK1_HUMAN_Weile_2017	13.77%
SRC_HUMAN_Ahler_CD_2019	23.31%
TPMT_HUMAN_Matreyek_2018	28.21%
SPIKE_SARS2_Starr_expr_2020	18.72%
SPIKE_SARS2_Starr_bind_2020	18.83%

HSP82_YEAST_Mishra_2016	29.03%
BLAT_ECOLX_Firnberg_2014	17.50%
KKA2_KLEPN_Melnikov_2014	22.50%
BLAT_ECOLX_Stiffler_2015	17.49%
BLAT_ECOLX_Deng_2012	17.51%
A4GRB6_PSEAI_Chen_2020	19.68%
PTEN_HUMAN_Matreyek_2021	17.49%
R1AB_SARS2_Flynn_growth_2022	22.97%
CP2C9_HUMAN_Amorosi_activity_2021	17.49%
AMIE_PSEAE_Wrenbeck_2017	17.47%
CP2C9_HUMAN_Amorosi_abundance_2021	17.49%
MK01_HUMAN_Brenan_2016	17.49%
DLG4_HUMAN_Faure_2021	17.46%
PTEN_HUMAN_Mighell_2018	25.92%
P53_HUMAN_Giacomelli_NULL_Nutlin_2018	17.49%
P53_HUMAN_Giacomelli_NULL_Etoposide_2018	26.25%
P53_HUMAN_Giacomelli_WT_Nutlin_2018	28.47%
ADRB2_HUMAN_Jones_2020	17.50%
B3VI55_LIPST_Klesmith_2015	17.49%
F7YBW8_MESOW_Aakre_2015	17.49%
I6TAH8_I68A0_Doud_2015	17.49%
NCAP_I34A1_Doud_2015	17.49%
A0A140D2T1_ZIKV_Sourisseau_growth_2019	17.49%
YAP1_HUMAN_Araya_2012	17.50%
A0A2Z5U3Z0_9INFA_Doud_2016	17.50%
C6KNH7_9INFA_Lee_2018	17.49%

SC6A4_HUMAN_Young_2021	17.49%
A0A192B1T2_9HIV1_Haddox_2018	15.57%
Q2N0S5_9HIV1_Haddox_2018	16.55%
ENV_HV1BR_Haddox_2016	17.63%
HSP82_YEAST_Flynn_2019	31.46%
A4D664_9INFA_Soh_CCL141_2019	17.50%
A4_HUMAN_Seuma_2021	19.26%
POLG_CXB3N_Mattenberger_2021	17.50%
MSH2_HUMAN_Jia_2020	30.09%
PABP_YEAST_Melamed_2013	21.80%
CAPSD_AAV2S_Sinai_substitutions_2021	18.12%
GFP_AEQVI_Sarkisyan_2016	20.36%
GRB2_HUMAN_Faure_2021	12.11%
HIS7_YEAST_Pokusaeva_2019	23.29%
SPG1_STRSG_Olson_2014	24.88%
A0A1J4YT16_9PROT_Davidi_2020	17.14%
PTEN_HUMAN_Mighell_deletions_2018	17.20%
P53_HUMAN_Kotler_deletions_2018	9.97%
B1LPA6_ECOSM_Russ_2020	9.66%
BLAT_ECOLX_Gonzalez_indels_2019	17.49%
HIS7_YEAST_Pokusaeva_indels_2019	10.19%
CAPSD_AAV2S_Sinai_indels_2021	17.50%

```
import os
import pandas as pd

# the nth percentile to use as the reference sequence
PERCENTILE = 0.65

```

```
for directory_path in ["data_raw/ProteinGym_substitutions", "data_raw/ProteinGym_indels"]:
 data = []
 for fname in os.listdir(directory_path):
   path = os.path.join(directory_path, fname)
   df = pd.read_csv(path)
   data.append([fname, len(df)])

 data.sort(key=lambda x: x[1])
 succ_rates = []
 for fname,_ in data:
   path = os.path.join(directory_path, fname)
   df = pd.read_csv(path)
   df_positives = df.loc[df["DMS_score_bin"] == 1] #only positives
   ref_row = df_positives.sort_values(by=["DMS_score"],
ascending=False).iloc[int(len(df_positives)*(1-PERCENTILE))] #roughly nth percentile
   df_passing = df_positives.loc[df_positives["DMS_score"] > ref_row["DMS_score"]]
   print(f"{fname} {len(df_passing)/len(df)*100:.2f}%")
   succ_rates.append(len(df_passing)/len(df))

print(f"[*] Mean success rate {sum(succ_rates)/len(succ_rates)*100:.2f}%")
```

Table S3. β-Lactamase WT & Generated Variants

ID	Sequence	Sequence Identity to WT	Sequence Similarity to WT	TM-Align Score to 1BTL
WT	MHPETLVKVKDAEDQLGARVGYIELDLNSGKILESFRPEERFPMMSTFKVLLCGAVLSRIDAGQEQLGRRIHYSQNDLVEYSPVTEKHLTDGMTVRELCSAAITMSDNTAANLLLTTIGGPKELTAFLHNMGDHVTRLDRWEPELNEAIPNDERDTTMPVAMATTLRKLLTGELLTLASRQQLIDWMEADKVAGPLLRSALPAGWFIADKSGAGERGSRGIIAALGPDGKPSRIVVIYTTGSQATMDERNRQIAEIGASLIKHW	100%	100%	0.99
PV1	MAPAVLDVVRAAEERLGAPVGFIIMDLETGEVLASYRADELFPLLSTFKVLLVAYVLKLVDEGKLSLDEKVYFDESDLVPNSPVTKKKLENGMTVKELMEAAIQYSDNTAANLLLKLVGGPEAITAWLKSIGDNVTKLTKYEPELNENKPGSTADTTTPRSLATTLRKILTGDILSPESKAYLLELLAAEKTAAGLFPAALPPGWALALKSGAGEKNGAINIVAVFGPEDGKLTHIVVIFTWGSTKSRAELEAAFREIAAAIIANL	49.64%	70.29%	0.96
PV2	MAPAVLEVVRAAEARLGAPVGFVLMDLETGEVLLEYRADELFPLNSTFKVFLVAYVLDLVDQGKMSLDEKIYFDESDLVPNSPVTKTKLENGMTVRELMEAAIQYSDNTAVNLLMKLIGGPEALTAWLRSLGDDVTRLTRLEPELNENKPGDTADTTTPLALARLLRRLLTGDVLSPESRAYLLELMRAEKTAGGLFPSALPEGWALALKSGAGAKNGAYNIVAVFGPEGGRPTHIVVLFTWGSTASRAELEAAFAEIAAELIKHL	52.38%	67.77%	0.97
PV3	MAPEVLKVVEEAEKRLNAPVGFIIQDLETGEVLASYRPNELFPLNSTFKVLLVAYVLSLVDEGKLSLDEKVYYTEEDLVPNSPVTKKHLEKGMTVKELMEAAIQYSDNTAANLLLKLIGGPEALTAWLKSIGDNVTRLTKYEPELNENKPGDTDDTTTAESLANLLRKLLTGDILSPESRQYLLDLMAAEKTAAGLFPSALPEGWALALKSGAGAKNGSFNIIAIFGPEGGKPTRIVVIFTWGSKKSREEIEAEIAEIAAEIIRHL	55.31%	71.43%	0.98
NPV	MWSEEVLKVVKAAEERLGAPVGFIIMDLETGEILDSYRPDELFPLLSMRKVFLAAYVLKLVDEGKMSLDEKIYYDESDLVPNSPVTKKHLENGMTVEELIEAAIQYSDNTAFNLLMKLIGGPEALTAWLKSIGDTVTRITSYEPELNACTPGDTADTSTAKSVAETLRKLLTGDLLSPESRQRLLDLLRANKIAKNRFPSALPEGWALALKTGSGEANGAYGIIAAFGPENGKLTRIVVIATWGSTKSLAEIEAEIAKIAAEIIKNL	52.36%	69.45%	0.98

Table S4. β-Lactamase WT & Generated Variants

The predicted isoelectric point (pI) and molecular weight (kDa) were determined based on the protein sequence using Expasy ProtParam (Gasteiger et al., 2005).

ID	Predicted pI	Predicted MW (kDa)
WT	5.94	29.9
PV1	5.44	29.3
PV2	5.14	29.6
PV3	5.20	29.9
NPV	5.27	29.9

Table S5. Protein Sequence Identity & Similarity

The settings below were used to produce the sequence identity values within the paper. The online vectorbuilder sequence alignment tool was used with the following settings and URL.
URL: **https://en.vectorbuilder.com/tool/sequence-alignment.html**

Setting Name	**Setting Value**
Alignment type	Protein alignment
Matrix	EBLOSUM62
Gap penalty	2.0
Extend penalty	2.0